Listening. Learning. Leading.®

Research Report
ETS RR–14-19

Effect of Item Response Theory (IRT) Model Selection on Testlet-Based Test Equating

Yi Cao

Ru Lu

Wei Tao

December 2014

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Effect of Item Response Theory (IRT) Model Selection on Testlet-Based Test Equating

Yi Cao,[1] Ru Lu,[1] & Wei Tao[2]

1 Educational Testing Service, Princeton, NJ
2 ACT, Iowa City, IA

The local item independence assumption underlying traditional item response theory (IRT) models is often not met for tests composed of testlets. There are 3 major approaches to addressing this issue: (a) ignore the violation and use a dichotomous IRT model (e.g., the 2-parameter logistic [2PL] model), (b) combine the interdependent items to form a polytomous item and apply a polytomous IRT model (e.g., the graded response model [GRM]), and (c) apply a model that explicitly takes into account the dependence at the item level (e.g., the testlet response theory [TRT] model). In this study, a simulation was conducted to compare the performance of these 3 approaches on number-correct score equating when degrees of testlet effect were manipulated. The traditional equipercentile method was used as an evaluation baseline. The results show that the 2PL and the TRT approaches produce comparable results that more closely agree with the results of the equipercentile method than the GRM does. And the number-correct equating using the 2PL is robust to the violation of local item independence.

Keywords Testlet; local item dependence; dichotomous item response model; polytomous item response model; the testlet response model; true score equating; observed score equating

doi:10.1002/ets2.12017

In the current practice of educational measurement, it is not uncommon for a standardized test to consist of testlets. A *testlet* is defined as an aggregation of items on a single theme (Wainer & Kiely, 1987). As the testlet items (only multiple-choice items are considered in this study) are designed to be assembled and administrated together under a common stimulus, items within a testlet often tend to violate the item response theory (IRT) assumption of local item independence and display some degree of local item dependence (LID), a testlet effect. Although an abundance of studies has examined the impact of testlet-caused LID on parameter recovery and proposed different approaches to accommodate LID, little research in the literature has focused on the effect of different approaches to handling LID on IRT-based number-correct score equating.

Three major approaches in operational practice are used to handle the LID caused by testlets. One approach is to ignore LID and treat the testlet items as discrete and locally independent and then apply unidimensional dichotomous IRT models, such as one- (1PL), two- (2PL), or three-parameter logistic (3PL) models. The second approach is to combine all interdependent items within a testlet into a single polytomous item and apply unidimensional polytomous IRT models, such as the graded response model (GRM), the generalized partial credit model, or the nominal response model. The third approach retains item-level information by explicitly modeling LID due to testlet effects under a multidimensional IRT framework. The bifactor model (Gibbons & Hedeker, 1992), the testlet response theory (TRT) model (Bradlow, Wainer, & Wang, 1999; Wainer, Bradlow, & Du, 2000; Wainer, Bradlow, & Wang, 2007; Wainer & Wang, 2000), and its modified version (Li, Bolt, & Fu, 2006) belong to this approach.

Results under the first approach (the dichotomous IRT approach) simply indicate the robustness of traditional IRT models to LID. Plenty of research has shown that the dichotomous IRT approach could lead to misestimation of item parameters and test reliability (Keller, Swaminathan, & Sireci, 2003; Lawrence, 1995; Sireci, Thissen, & Wainer, 1991; Zenisky, Hambleton, & Sireci, 2002). However, very few studies have focused on examining the impact of LID on number-correct score equating when traditional IRT models are applied.

The second approach (the polytomous IRT approach) is easy in interpretation and implementation, but it suffers the problem of losing response pattern information due to combining items (Sireci et al., 1991; Zenisky et al., 2002).

Corresponding author: Y. Cao, E-mail: ycao@ets.org

Lee, Kolen, Frisbie, and Ankenmann (2001) used real data to compare the performance of dichotomous and polytomous models on number-correct score equating in testlet-based tests. They found that treating a testlet as a polytomous item and using it as the unit of analysis was more effective in equating than ignoring LID and using the traditional 3PL model.

The third approach retains item-level information but requires building a more complex model. Two studies focused on IRT linking/equating using this approach: one related to scale transformation and the other related to number-correct score equating. Li, Bolt, and Fu (2005) developed scale transformation procedures to extend the traditional test characteristic curve method to the 2PL ogive TRT model under the nonequivalent groups with anchor test equating design. They investigated the effectiveness of their proposed method via simulation and found that when LID due to testlet effects was present, their proposed method better recovered the linking coefficients compared to the traditional IRT method. In their study, they focused only on the TRT scale transformation method, not on the TRT number-correct score equating such as the true score equating (TSE) and observed score equating (OSE) methods. Tao and Cao (2012) proposed procedures to conduct IRT TSE and OSE with the modified TRT model and compared the performance of the traditional 3PL and the TRT model on number-correct score conversions when various degrees of LID were present. Their results showed that when LID was at a moderate or high level, the TRT model yielded more accurate equating results compared to those using the traditional 3PL. However, their study did not include polytomous IRT models as an approach to accommodating LID and only compared the number-correct score equating results among the dichotomous IRT and TRT approaches.

Many testing programs use IRT-based methods to conduct equating to place number-correct scores from different forms onto a common scale. The proper selection and application of IRT models in handling LID have an influence on item and ability parameter estimation in testlet-based tests, which consequently could have a practical impact on the accuracy of IRT number-correct score equating results. Therefore, the main purpose of this study is to conduct a simulation study to compare the number-correct equating results in terms of the raw-to-raw conversions among the three approaches mentioned before—the dichotomous IRT, the polytomous IRT, and the TRT approaches—when various degrees of LID due to testlets are present. The first approach allows one to investigate the impact of LID on number-correct equating results when a selected dichotomous IRT model is applied.

In the rest of the article, all the models used in this study are first introduced with selection reasons, followed by an illustration of conducting TSE and OSE with the TRT model. Then, the simulation design, results, and conclusions are presented.

Item Response Theory (IRT) Models

The Two-Parameter Logistic (2PL) Model

The 2PL was selected to represent the dichotomous IRT approach. Under the 2PL, the probability of examinee j correctly answers item i can be expressed as:

$$P\left(X_{ij}=1|\theta_j\right)=\frac{1}{1+\mathrm{e}^{-a_i\left(\theta_j-b_i\right)}}, \tag{1}$$

where θ_j is the primary trait designed to be measured by the test for examinee j, a_i is the discrimination parameter for item i, and b_i is the difficulty parameter for item i.

The reason for selecting the 2PL over the 3PL is as follows. Previous studies (Wainer & Wang, 2000; Wainer et al., 2000) have shown that when the traditional 3PL was applied to the situation where LID was present, the c-parameter was often misestimated. Meanwhile, as Kolen and Brennan (2004) pointed out, when using the 3PL in TSE, the c-parameters posed a floor effect on base form true score equivalents, which in turn required a linear interpolation at the lower end of the score scale. Tao and Cao (2012) further suggested that the inaccuracy of the c-parameter estimation and the inclusion of the linear interpolation might account for the worse performance of the TSE compared to the OSE. In order to make the TSE and OSE more comparable, the 2PL was selected for this study.

The Graded Response Model (GRM)

The GRM was selected to represent the polytomous IRT approach to accommodating LID due to testlet effects. The GRM is appropriate to use when item responses can be characterized as ordered categorical responses. The testlet item scores would have an ordered quality if they related to the extent of the completeness of an examinee's reasoning process within

a specific testlet. The more items within a testlet that an examinee could answer correctly, the more extensive his or her reasoning process would be. In this sense, using the GRM in testlet-based equating is an appropriate application. Alternatively, the generalized partial credit model is another valid choice for ordered categorical responses. Because neither model consistently exhibits superiority over the other based on the existing literature (Cao, Yin, & Gao, 2007; Lee et al., 2001; Tang & Eignor, 1997), the selection of GRM is arbitrary in this study.The GRM directly models the cumulative category response function. Under the GRM, the probability of examinee j earning a score on item i at or above category k can be expressed as:

$$P^*_{ijk}\left(X_{ijk} \geq x_k|\theta_j\right) = \begin{cases} 1 & k = 1 \\ \frac{1}{1+e^{-a_i\left(\theta_j - b_{ik}\right)}} & 2 \leq k \leq K \\ 0 & k > K \end{cases}, \tag{2}$$

where category $k = 1, 2, \ldots, K$, a_i is the item slope parameter. All the category characteristic curves for a given item share the same a_i. b_{ik} is the between category threshold parameter of category k in item i, whose value represents the point on the θ continuum where individuals have a 50% chance of responding at or above category k. Once the $P^*_{ijk}\left(\theta_j\right)$ is estimated, the actual category response function can be computed using the following equation:

$$P\left(X_{ijk} = x_k|\theta_j\right) = P^*_{ijk}\left(\theta_j\right) - P^*_{ij(k+1)}\left(\theta_j\right). \tag{3}$$

It represents the probability of examinee j responding to a particular category k.

The Testlet Response Theory (TRT) Model

The TRT models were introduced in a series of papers (Bradlow et al., 1999; Wainer & Wang, 2000; Wainer et al., 2000; Wainer et al., 2007). Li et al. (2006) pointed out that the TRT model assumes that items that discriminate well on the primary trait also discriminate well on the testlet traits, when the opposite might seem more reasonable in practice. Despite its limitation, the TRT model has received a substantial amount of attention and is still predominantly used in the recent literature on modeling LID due to testlet effects. Meanwhile, DeMars (2006) found evidence "favoring the use of the more parsimonious testlet-effects model over the bifactor model" (p. 166) when LID is present. Therefore, the TRT model is selected to represent the multidimensional IRT approach to accommodating LID due to testlet effects.

The two-parameter version of the TRT model can be expressed as:

$$P\left(X_{ij} = 1|\theta_j, \gamma_{d(i)j}\right) = \frac{1}{1 + e^{-a_i\left(\theta_j - b_i - \gamma_{d(i)j}\right)}}, \tag{4}$$

where $d(i)$ denotes a testlet containing item i. θ_j, a_i, and b_i have the same interpretations as in the traditional 2PL model. The $\gamma_{d(i)j}$ is referred to as the random testlet effect, which represents an interaction between testlet $d(i)$ and examinee j's ability on that testlet. Items within the same testlet have the same testlet effect. It can be further interpreted as the examinee's standing on a testlet-specific trait, independent of the primary trait θ_j. Thus, for a test containing D testlets, the TRT model has $D + 1$ dimension: one primary trait plus D testlet-specific traits. The model assumes that $\gamma_{d(i)j}$ follows a normal distribution as $\gamma_{d(i)j} \sim N\left(0,\ \sigma^2_{\gamma_{d(i)j}}\right)$. The magnitude of the testlet effect is reflected by $\sigma^2_{\gamma_{d(i)j}}$. The larger the $\sigma^2_{\gamma_{d(i)j}}$ is, the higher degree of LID among items within a testlet and the larger the testlet effect will be. If there is no testlet effect ($\sigma^2_{\gamma_{d(i)j}} = 0$), the TRT model is reduced to the traditional 2PL model.

Number-Correct Score Equating With the Testlet Response Theory (TRT) Model

IRT TSE and OSE are the two methods that can be used to put number-correct scores of a new form onto a reference form scale. IRT TSE and OSE with the traditional 2PL and the GRM are well established and documented in Kolen and Brennan (2004). Tao and Cao (2012) proposed and explained their procedures to conduct IRT TSE and OSE with the three-parameter version of the modified TRT model. The current study adopts their procedures and tailors them to conduct IRT TSE and OSE with the two-parameter version of the TRT model.

The Testlet Response Theory (TRT) True Score Equating (TSE)

The TRT TSE follows the similar three-step process that is used in traditional IRT TSE. In this process, the first step is to specify the number-correct true score (τ) on a new form. Then, find the θ corresponding to that true score. Last, find the true score on a reference form associated with that same θ. The difference between the TRT model and the traditional IRT model is that there are two θ s (i.e., a primary trait and a testlet-specific trait) instead of one determining the probability of a correct response to an item in the TRT model. Tao and Cao (2012) proposed to equate the new and reference forms only through the primary trait by integrating out the testlet-specific traits. The testlet-specific trait is usually not designed to be measured and cannot generalize across contexts. For this reason, the testlet-specific trait is often regarded as a nuisance trait and only the primary trait is of interest in the TRT model.

For the TRT TSE, the first and last steps are straightforward. The second step, to find the primary trait θ_j, is a critical step, and the Newton-Raphson method is applied to do so (Kolen & Brennan, 2004; Tao & Cao, 2012). The Newton-Raphson method is an iterative process used for finding successively better approximations to the root of a nonlinear function. It is implemented as follows: Begin with a function that is set to 0. Given that function func(θ_j) defined over the variable θ_j and its first derivative with respect to θ_j as func$'(\theta_j)$, an initial value is chosen for θ_j, which is referred to as θ_j^-. A new value for θ_j, θ_j^+, is calculated as:

$$\theta_j^+ = \theta_j^- - \frac{\text{func}\left(\theta_j\right)}{\text{func}'\left(\theta_j\right)}. \tag{5}$$

Typically, θ_j^+ will be closer to the root of the function func(θ_j) than θ_j^-. The new value is then redefined as θ_j^-, and the process is repeated until θ_j^+ and θ_j^- are equal (i.e., func(θ_j) is close to 0) at a specified level of precision.

More specifically, in the TRT TSE, func(θ_j) and func$'(\theta_j)$ are defined as:

$$\text{func}\left(\theta_j\right) = \tau - \sum_i P\left(X_{ij} = 1|\theta_j\right), \tag{6}$$

$$\text{func}'\left(\theta_j\right) = -\sum_i \frac{\partial P\left(X_{ij} = 1|\theta_j\right)}{\partial \theta_j}, \tag{7}$$

where τ is the number-correct true score on a new form whose equivalent is to be found, and $\sum_i P\left(X_{ij} = 1|\theta_j\right)$ is the test characteristic curve, which is the summation of the marginalized item response functions of the primary trait θ_j over all items on a new form. This marginalized item response function of the primary trait θ_j in the TRT model can be expressed as:

$$P\left(X_{ij} = 1|\theta_j\right) = \int_{\gamma_{d(i)j}} P\left(X_{ij} = 1|\theta_j, \gamma_{d(i)j}\right)\phi\left(\gamma_{d(i)j}\right) d\gamma_{d(i)j}, \tag{8}$$

where $\phi(\gamma_{d(i)j})$ is the density of $\gamma_{d(i)j}$, which is assumed to follow a normal distribution. A discrete distribution on a finite number of equally spaced points can be used to approximate the integral,

$$P\left(X_{ij} = 1|\theta_j\right) = \sum_t P\left(X_{ij} = 1|\theta_j, \varphi_{d(i)jt}\right) A\left(\varphi_{d(i)jt}\right), \tag{9}$$

where $\varphi_{d(i)jt}$ and $A(\varphi_{d(i)jt})$ represent the node and weight of $\gamma_{d(i)j}$ at quadrature point t. Forty-one quadrature points were used for the testlet-specific traits in this study.

Then, the first derivative of $P(X_{ij} = 1|\theta_j)$ with respect to the primary trait θ_j can be expressed as:

$$\frac{\partial P\left(X_{ij} = 1|\theta_j\right)}{\partial \theta_j} = \sum_t a_i \left[1 - P\left(X_{ij} = 1|\theta_j, \varphi_{d(i)jt}\right)\right] P\left(X_{ij} = 1|\theta_j, \varphi_{d(i)jt}\right) A\left(\varphi_{d(i)jt}\right). \tag{10}$$

Substitute Equations 9 and 10 into Equations 6 and 7 and the resulting expressions for func(θ_j) and func$'(\theta_j)$ are then substituted into Equation 5 to solve for θ_j.

The Testlet Response Theory (TRT) Observed Score Equating (OSE)

The TRT OSE follows the similar process as the traditional IRT OSE. In this process, the IRT model is used to produce estimated distributions of observed number-correct scores on new and reference forms, which are then equated using an equipercentile method. Usually, a compound binomial distribution is assumed and a recursion formula (Kolen & Brennan, 2004; Lord & Wingersky, 1984) is used to first find the observed score distribution for a given θ. Then these distributions are accumulated over the whole θ scale to produce an estimated number-correct score distribution of the form. For example, an estimated number-correct score distribution of the new form can be expressed as:

$$f(x) = \int_{\theta} f(x|\theta)\,\phi(\theta)\,d\theta. \tag{11}$$

For the TRT OSE, one more step is applied to first find the observed score distribution for a given primary trait θ_j by integrating out the testlet-specific traits:

$$f\left(x|\theta_j\right) = \int_{\gamma_{d(i)j}} f\left(x|\theta_j, \gamma_{d(i)j}\right)\phi\left(\gamma_{d(i)j}\right) d\gamma_{d(i)j}, \tag{12}$$

which then is substituted into Equation 11 to solve $f(x)$,

$$f(x) = \int_{\theta_j} f\left(x|\theta_j\right)\phi\left(\theta_j\right) d\theta_j = \int_{\theta_j}\int_{\gamma_{d(i)j}} f\left(x|\theta_j, \gamma_{d(i)j}\right)\phi\left(\gamma_{d(i)j}\right) d\gamma_{d(i)j}\phi\left(\theta_j\right) d\theta_j. \tag{13}$$

A discrete distribution on a finite number of equally spaced points can be used to approximate the integral,

$$f(x) = \sum_{s}\sum_{t} f\left(x|\varphi_{js}, \varphi_{d(i)jt}\right) A\left(\varphi_{d(i)jt}\right) A\left(\varphi_{js}\right), \tag{14}$$

where φ_{js} and $A(\varphi_{js})$ represent the node and weight of the primary θ_j at quadrature point s; $\varphi_{d(i)jt}$ and $A(\varphi_{d(i)jt})$ represent the node and weight of $\gamma_{d(i)j}$ at quadrature point t. In this study, 41 quadrature points were used for both θ_j and $\gamma_{d(i)j}$.

The same procedures are used to generate an estimated number-correct score distribution of the reference form $f(y)$. Then, an equipercentile method is applied for equating.

Method

Equating Design

This simulation study employed the random groups equating design and the IRT-based equating methods. In the random groups design, two samples of examinees were randomly selected from the same population with one sample taking the new form and the other sample taking the reference form. Three IRT models—the 2PL model, the GRM, and the TRT model—were selected to represent three different approaches to accommodating LID due to testlet effects and to equate testlet-based tests. Separate calibrations with the same scaling convention (i.e., mean of 0 and standard deviation of 1 for the ability prior) were conducted to put the IRT parameters for the new form onto the reference form scale. No further scale transformation was needed. Then the IRT TSE and OSE methods were used to produce equated number-correct scores on the new form. In total, for each new-to-reference form equating, six raw-to-raw conversion tables were generated (crossing three IRT models by two equating methods). They were abbreviated as 2PL TSE, 2PL OSE, GRM TSE, GRM OSE, TRT TSE, and TRT OSE.

Data Generation

Four tests with varying degrees of LID were simulated in this study. They were all designed to measure a single latent trait and were specified to reflect reasonable configurations for large-scale assessments. Each test had a total of 40 multiple-choice items, composed of 10 discrete items and 6 testlets with 5 items per testlet. The first test, T0, consisted of all locally independent multiple-choice items. The second test, TL, was composed of both discrete and low LID items; the third test, TM, of both discrete and moderate LID items; and the fourth test, TH, of both discrete and high LID items. Based on previous studies (Bradlow et al., 1999; DeMars, 2006; Li et al., 2006; Zu & Liu, 2010), the degree of LID due to testlet effects,

indexed by $\sigma^2_{\gamma_{d(i)j}}$ in Equation 4, was set to four levels representing four tests: zero, low (uniformly distributed between [0.1 – 0.5]), moderate (uniformly distributed between [0.6 – 1.0]), and high (uniformly distributed between [1.1 – 1.5]).

Two test forms per test (the new and the reference form) were simulated for equating. The item difficulty parameters for the reference form were randomly drawn from a standard normal distribution with the range from −2.5 to 2.5. The item discrimination parameters were sampled from a log-normal distribution with mean of 0 and standard deviation of 0.5, constrained to 0.5 – 2.5. The item parameters for the new form were drawn from the same distributions as the reference form, except for the item difficulty parameters. As the purpose of equating is to adjust difficulty differences across forms statistically, the item difficulty parameter distribution in the new form was intentionally randomly drawn from a normal distribution of (0.1, 1) to indicate that the new form was slightly more difficult than the reference form. The item parameters and the associated testlet effects used to simulate the response data are presented in Table 1.

The item responses of examinees taking the new and the reference forms were generated separately. For each sample, 2,000 examinees' responses were created. For each examinee, a primary trait was randomly drawn from a standard normal distribution and six testlet-specific traits were independently drawn from normal distributions with mean of 0 and variances as specified in Table 1. Based on the simulated item parameters, the primary trait and the appropriate testlet-specific trait, the probability of each examinee's correct response to each item was calculated using the TRT model in Equation 4. Then this probability was compared to a random draw from a uniform distribution between 0 and 1: If the random draw was less than the probability, the response was coded as correct (i.e., 1); otherwise, as incorrect (i.e., 0). The individual item scores in each testlet were summed, and the sum was treated as a single item score when using the GRM. This data generation process was repeated 50 times for each new and reference form of each test.

The program SAS was used for data generation. BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003), PARSCALE (Muraki & Bock, 2003), and SCORIGHT (Wang, Bradlow, & Wainer, 2005) were used to calibrate the response data by the 2PL model, the GRM, and the TRT model, respectively. The program POLYEQUATE (Kolen, 2004) was used to conduct TSEs and OSEs with the 2PL and the GRM. An SAS program was written to conduct the TRT TSEs and OSEs.

Evaluation Criteria

The classical equipercentile method was used as the baseline for evaluation because it only employed total test scores and was not influenced by the violation of LID. The equipercentile equating function is developed by identifying scores on a new form that have the same percentile ranks as scores on a reference form (Kolen & Brennan, 2004). For each of the four tests, one equipercentile equating was conducted to yield a population conversion (based on a population of 100,000 examinees [2,000 examinees × 50 replications]) using RAGE-RGEQUATE (Zeng, Kolen, Hanson, Cui, & Chien, 2005).

The equating bias, standard error of equating (SEE), and root mean squared error (RMSE) were used to evaluate the difference between an IRT conversion and the population equipercentile conversion at each raw score point. In addition, weighted averages of these indices across all score points were computed to evaluate an overall discrepancy at the test level.

Equating bias is an index of systematic error of equating, and the conditional bias at each score point x is defined as:

$$\text{bias}(x) = \frac{1}{R}\sum_{r=1}^{R}\left[\hat{e}_r(x)\right] - e(x) = \bar{\hat{e}}(x) - e(x), \tag{15}$$

where $\hat{e}_r(x)$ is the estimated reference form equivalent of score point x on the new form in the rth replication, and $e(x)$ is the reference form equivalent of score point x in the population conversion. R is the total number of replications (equal to 50 in this study). $\bar{\hat{e}}(x)$ is the average of $\hat{e}_r(x)$ over the R replications. Then the individual biases could be aggregated to reach an overall measure of systematic errors across all score points, which is the weighted average of bias, $\sqrt{\sum_x f(x)\text{bias}^2(x)}$. $f(x)$ is the raw proportion of examinees at score point x on the new form. And the squared value is taken to ensure the positive and negative biases at various score points will not be canceled out.

SEE is an index of random sampling error in equating, and it is defined as:

$$\text{SEE}(x) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left[\hat{e}_r(x) - \bar{\hat{e}}(x)\right]^2}. \tag{16}$$

Table 1 Item Parameters for the Reference Form and New Form

		Reference form						New form					
				σ_γ^2						σ_γ^2			
Item	Testlet	*a*	*b*	T0	TL	TM	TH	*a*	*b*	T0	TL	TM	TH
1		1.7	−0.2	0.0	0.0	0.0	0.0	1.2	−0.1	0.0	0.0	0.0	0.0
2		1.1	−0.8	0.0	0.0	0.0	0.0	0.9	0.6	0.0	0.0	0.0	0.0
3		2.1	1.5	0.0	0.0	0.0	0.0	0.5	0.7	0.0	0.0	0.0	0.0
4		0.9	0.4	0.0	0.0	0.0	0.0	0.9	0.5	0.0	0.0	0.0	0.0
5		0.9	1.3	0.0	0.0	0.0	0.0	1.2	0.1	0.0	0.0	0.0	0.0
6		0.5	0.7	0.0	0.0	0.0	0.0	2.0	0.4	0.0	0.0	0.0	0.0
7		1.3	0.3	0.0	0.0	0.0	0.0	1.0	0.3	0.0	0.0	0.0	0.0
8		0.8	0.0	0.0	0.0	0.0	0.0	0.7	0.6	0.0	0.0	0.0	0.0
9		0.6	−0.2	0.0	0.0	0.0	0.0	1.2	−1.0	0.0	0.0	0.0	0.0
10		2.1	−1.9	0.0	0.0	0.0	0.0	0.7	−1.0	0.0	0.0	0.0	0.0
11	1	0.8	−0.7	0.0	0.3	0.6	1.4	0.8	0.6	0.0	0.3	0.8	1.5
12	1	1.5	2.1	0.0	0.3	0.6	1.4	1.4	−0.7	0.0	0.3	0.8	1.5
13	1	0.6	−1.9	0.0	0.3	0.6	1.4	0.7	−0.4	0.0	0.3	0.8	1.5
14	1	0.8	−1.0	0.0	0.3	0.6	1.4	0.6	0.7	0.0	0.3	0.8	1.5
15	1	1.0	0.0	0.0	0.3	0.6	1.4	0.9	0.5	0.0	0.3	0.8	1.5
16	2	1.2	0.4	0.0	0.5	0.8	1.1	0.8	2.1	0.0	0.4	0.7	1.5
17	2	1.3	−0.7	0.0	0.5	0.8	1.1	1.3	−1.6	0.0	0.4	0.7	1.5
18	2	1.0	0.3	0.0	0.5	0.8	1.1	1.2	−2.3	0.0	0.4	0.7	1.5
19	2	0.8	−0.3	0.0	0.5	0.8	1.1	0.8	−0.9	0.0	0.4	0.7	1.5
20	2	0.5	−1.0	0.0	0.5	0.8	1.1	1.9	1.5	0.0	0.4	0.7	1.5
21	3	0.6	1.1	0.0	0.1	0.8	1.5	1.2	1.4	0.0	0.5	0.9	1.4
22	3	1.1	0.5	0.0	0.1	0.8	1.5	0.8	0.3	0.0	0.5	0.9	1.4
23	3	0.8	0.4	0.0	0.1	0.8	1.5	1.2	1.6	0.0	0.5	0.9	1.4
24	3	1.3	−1.7	0.0	0.1	0.8	1.5	1.6	−0.4	0.0	0.5	0.9	1.4
25	3	1.3	−0.4	0.0	0.1	0.8	1.5	0.7	0.4	0.0	0.5	0.9	1.4
26	4	1.1	0.0	0.0	0.2	0.9	1.5	1.1	−0.8	0.0	0.2	0.9	1.3
27	4	1.6	−0.9	0.0	0.2	0.9	1.5	1.1	−0.3	0.0	0.2	0.9	1.3
28	4	1.1	0.3	0.0	0.2	0.9	1.5	1.4	0.0	0.0	0.2	0.9	1.3
29	4	0.8	0.0	0.0	0.2	0.9	1.5	1.2	−1.0	0.0	0.2	0.9	1.3
30	4	1.1	−0.1	0.0	0.2	0.9	1.5	1.3	1.2	0.0	0.2	0.9	1.3
31	5	1.1	−0.2	0.0	0.5	1.0	1.4	2.5	−1.2	0.0	0.5	1.0	1.1
32	5	1.0	−1.0	0.0	0.5	1.0	1.4	1.0	2.0	0.0	0.5	1.0	1.1
33	5	0.8	0.4	0.0	0.5	1.0	1.4	1.4	−0.2	0.0	0.5	1.0	1.1
34	5	1.8	0.0	0.0	0.5	1.0	1.4	1.8	0.7	0.0	0.5	1.0	1.1
35	5	1.9	1.3	0.0	0.5	1.0	1.4	0.7	−1.6	0.0	0.5	1.0	1.1
36	6	1.4	0.9	0.0	0.3	0.8	1.3	0.8	−0.8	0.0	0.3	0.8	1.4
37	6	0.7	−1.2	0.0	0.3	0.8	1.3	0.9	0.7	0.0	0.3	0.8	1.4
38	6	1.0	−0.6	0.0	0.3	0.8	1.3	0.9	0.4	0.0	0.3	0.8	1.4
39	6	0.9	−0.4	0.0	0.3	0.8	1.3	1.1	0.8	0.0	0.3	0.8	1.4
40	6	0.5	−1.0	0.0	0.3	0.8	1.3	0.7	0.5	0.0	0.3	0.8	1.4
Mean		1.1	−0.1					1.1	0.1				
SD		0.4	0.9					0.4	1.0				

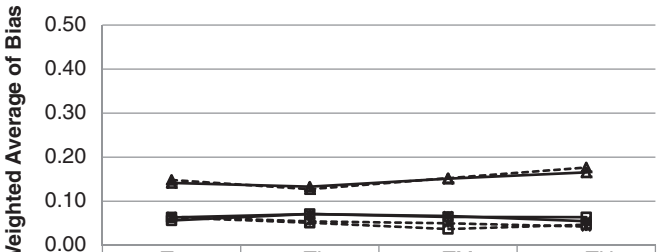
Note. T0 = zero local item dependence (LID) test; TL = low LID test; TM = moderate LID test; TH = high LID test.

Similarly, the weighted average of SEE could be expressed as $\sqrt{\sum_x f(x)\mathrm{SEE}^2(x)}$.

RMSE represents the combination of systematic and random errors and is defined as:

$$\mathrm{RMSE}(x) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left[\hat{e}_r(x) - e(x)\right]^2}. \quad (17)$$

Its corresponding weighted average could be expressed as $\sqrt{\sum_x f(x)\,\mathrm{RMSE}^2(x)}$. The relationship among bias, SEE, and RMSE could be proven to be $\mathrm{RMSE} = \sqrt{\mathrm{bias}^2 + \mathrm{SEE}^2}$.



	T0	TL	TM	TH
2PL-TSE	0.06	0.07	0.07	0.05
GRM-TSE	0.14	0.13	0.15	0.17
TRT-TSE	0.06	0.07	0.06	0.06
2PL-OSE	0.06	0.05	0.05	0.04
GRM-OSE	0.15	0.13	0.15	0.18
TRT-OSE	0.06	0.05	0.04	0.05

Figure 1 Weighted average of bias across different conditions. *Note.* T0 = zero local item dependence (LID) test; TL = low LID test; TM = moderate LID test; TH = high LID test; 2PL = two-parameter logistic; GRM = graded response model; OSE = observed score equating; TRT = testlet response theory; TSE = true score equating.

Results

Equating Bias

Figure 1 presents the weighted average of bias in both graphical and numeric formats. It compares the systematic discrepancy between each of the six IRT equating conversions and its corresponding equipercentile population conversion across four tests with varying degrees of LID. The weighted averages of bias are generally small, ranging from 0.04 to 0.18. The most obvious finding is that the GRM equating methods yield the largest biases among the three IRT-based methods; and the 2PL and TRT equating methods yield very comparable biases across all four test conditions. The TSEs and OSEs yield similar biases. The OSE method performs slightly but not significantly better than the TSE method does for the 2PL and TRT models. Furthermore, as the degree of LID increases, the weighted averages of bias yielded by the GRM equating methods increase slightly. However, the 2PL and TRT equating methods yield almost equivalent biases across four test conditions, which indicates that the 2PL equating methods are quite robust to the varying degree of LID caused by testlets.

To further explore the equating bias pattern across the IRT models, the equating methods, and the test conditions, Figure 2 is plotted to show the individual bias at each score point. It should be noted that the general patterns of the equating bias presented by the TSE and OSE methods are very similar, and thus, only the bias results produced by the true score method are used for interpretation and displayed in Figure 2. The results show that the biases at the two ends of the raw score scale are larger than the biases in the middle range of the scale. The biases yielded by the GRM TSE are the largest and have the largest fluctuations compared to those produced by the 2PL and TRT methods. Furthermore, when LID is not present (T0), the biases yielded by the 2PL and TRT methods are very similar, overlapping with each other. As LID increases from low to high (TL to TH), the difference in bias between the two methods gradually increases. This pattern is not apparent from the weighted average of bias as these individual discrepancies are balanced out when computing the weighted average of bias.

Standard Error of Equating (SEE)

Figure 3 shows the weighted average of SEE in both graphical and numeric formats. It compares the random sampling discrepancy across six IRT equating methods and four test conditions. The weighted averages of SEE range from 0.25 to 0.40, with the maximal SEEs occurring for the GRM equating methods in the moderate LID test condition and the minimums for the 2PL and TRT equating methods in the non-LID test. The GRM equating methods have slightly larger weighted averages of SEE than the 2PL and TRT equating methods, whereas the 2PL and TRT equating methods have very similar SEEs. The weighted averages of SEE produced by the TSEs and OSEs display almost no differences. Lastly, the weighted averages of SEE are the smallest for the test with zero LID (T0). They increase gradually as the degree of LID increases from zero to moderate LID then drop a little when the test has high LID.

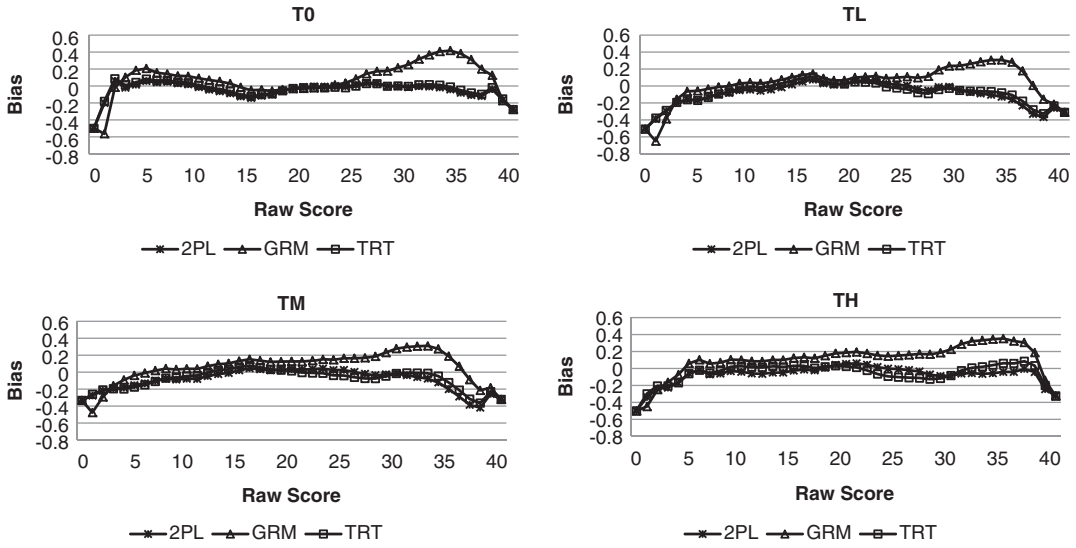
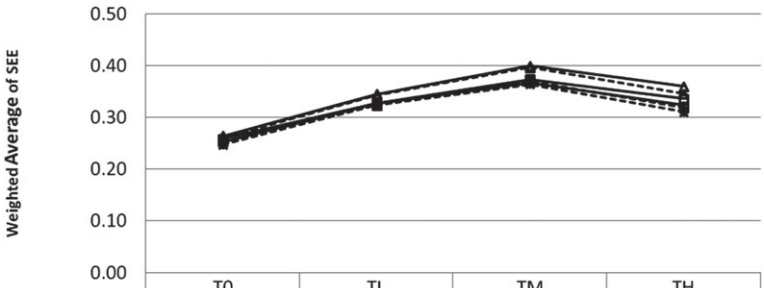


Figure 2 Equating bias for the true score equating (TSE) by test. *Note.* T0 = zero local item dependence (LID) test; TL = low LID test; TM = moderate LID test; TH = high LID test; 2PL = two-parameter logistic; GRM = graded response model; TRT = testlet response theory.



	T0	TL	TM	TH
2PL-TSE	0.25	0.33	0.37	0.32
GRM-TSE	0.26	0.35	0.40	0.36
TRT-TSE	0.26	0.33	0.37	0.34
2PL-OSE	0.25	0.32	0.36	0.31
GRM-OSE	0.26	0.34	0.40	0.35
TRT-OSE	0.25	0.32	0.37	0.32

Figure 3 Weighted average of standard error of equating (SEE) across different conditions. *Note.* T0 = zero local item dependence (LID) test; TL = low LID test; TM = moderate LID test; TH = high LID test; 2PL = two-parameter logistic; GRM = graded response model; OSE = observed score equating; TRT = testlet response theory; TSE = true score equating.

The individual SEEs at each score point across the IRT models, the equating methods, and the test conditions are plotted in Figure 4. As the general patterns of the SEEs presented by the TSE and OSE methods are similar (except for the minimum and maximum score points of 0 and 40, at which the TSE method sets them to be fixed, and thus, SEEs at these two points always equal zero for the TSE method), Figure 4 only displays the SEE results produced by the TSE method. It reveals that the GRM TSE method produces relatively larger SEE values, especially at the two ends of the score scale, whereas the SEE results yielded by the 2PL and TRT equating methods are comparable with slightly smaller SEEs by the 2PL method, especially when LID is moderate to high.

Root Mean Squared Error (RMSE)

Figure 5 shows the weighted average of RMSE in both graphical and numeric formats. It combines both the equating bias and SEE results and thus shows the overall discrepancy. Because the SEE values are much larger than the biases,

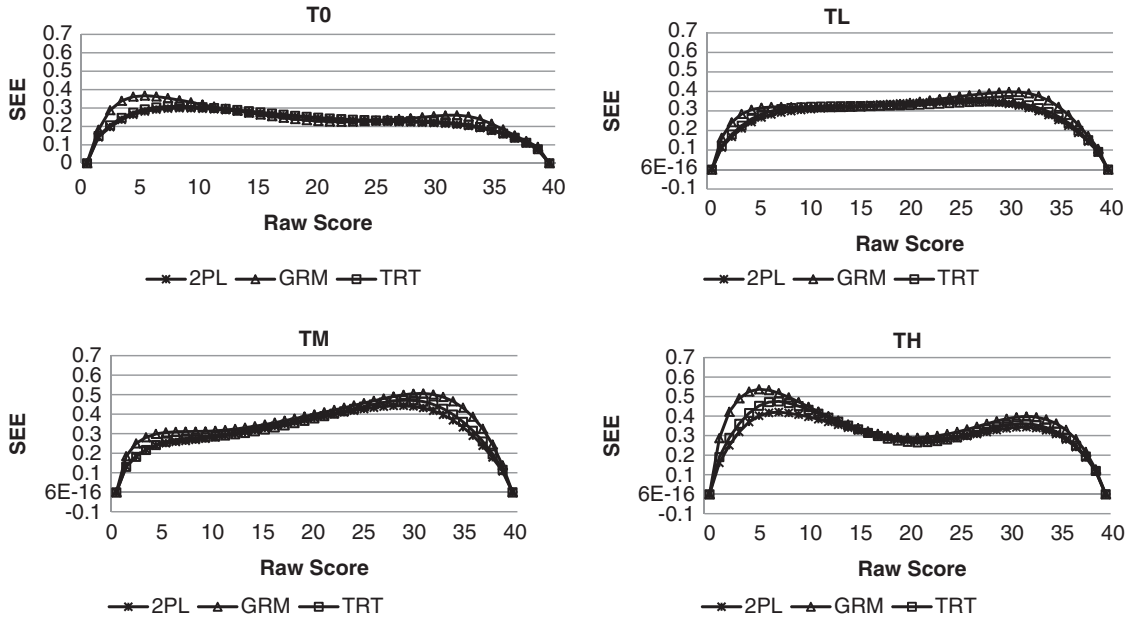
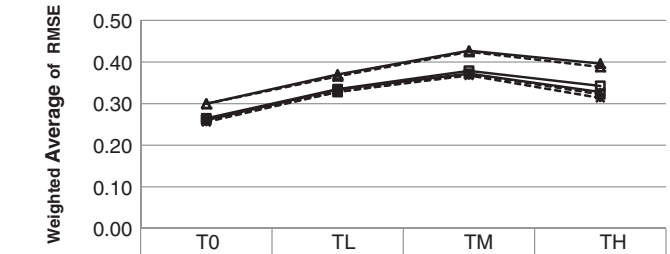


Figure 4 Standard error of equating (SEE) for the true score equating (TSE) by test. *Note.* T0 = zero local item dependence (LID) test; TL = low LID test; TM = moderate LID test; TH = high LID test; 2PL = two-parameter logistic; GRM = graded response model; TRT = testlet response theory.



	T0	TL	TM	TH
2PL-TSE	0.26	0.33	0.37	0.33
GRM-TSE	0.30	0.37	0.43	0.40
TRT-TSE	0.26	0.33	0.38	0.34
2PL-OSE	0.26	0.33	0.37	0.31
GRM-OSE	0.30	0.37	0.42	0.39
TRT-OSE	0.26	0.33	0.37	0.32

Figure 5 Weighted average of root mean squared error (RMSE) across different conditions. *Note.* T0 = zero local item dependence (LID) test; TL = low LID test; TM = moderate LID test; TH = high LID test; 2PL = two-parameter logistic; GRM = graded response model; OSE = observed score equating; TRT = testlet response theory; TSE = true score equating.

the patterns revealed in Figure 5 are primarily similar to those shown in Figure 3. First, the GRM equating methods consistently display the largest weighted averages of RMSE among the three IRT models across four test conditions. The magnitude of the RMSE discrepancy between the GRM and the 2PL/TRT equating methods are more obvious than shown in Figure 3 because the RMSE also incorporates the bias results. Meanwhile, the 2PL and TRT equating methods keep showing very comparable weighted averages of RMSE. Second, the TSEs and OSEs yield similar weighted averages of RMSE. Third, the weighted averages of RMSE are the smallest in the test without LID (T0). As the degree of LID increases from zero to moderate, the weighted averages of RMSE increase gradually, but they decrease when the test has high LID. However, further analyses show that the weighted averages of RMSE among the low, moderate, and high LID tests are not statistically significant.

Figure 6 shows the individual RMSE results at each score point produced by the TSE method. Again, the general patterns of the RMSE results are similar for the TSE and OSE methods except for the minimum and maximum score

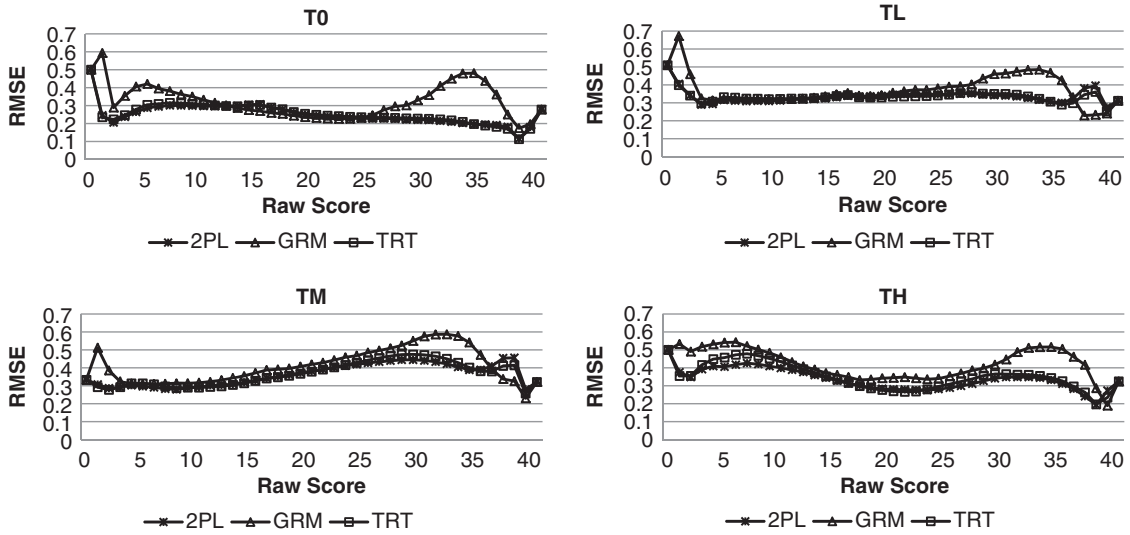


Figure 6 Root mean square error (RMSE) for the true score equating (TSE) by test. *Note.* T0 = zero local item dependence (LID) test; TL = low LID test; TM = moderate LID test; TH = high LID test; 2PL = two-parameter logistic; GRM = graded response model; TRT = testlet response theory.

points of 0 and 40. First, the GRM TSE method has the largest fluctuations across the score scale. Second, the 2PL and TRT equating methods yield very comparable RMSE values in the tests with zero or low LID. But as LID increases, the discrepancies of the RMSE values yielded by the 2PL and TRT methods at certain individual score points become more obvious.

Discussion and Conclusion

In this study, we examined the effect of selecting different IRT models on the raw-to-raw conversion tables for tests composed of varying degrees of LID caused by testlets. Of the three models selected in this study, the GRM and the TRT model were selected as a means to accommodate LID and the 2PL was selected to examine the impact of violating the local item independence assumption on the final conversion table.

Our three main conclusions are as follows. First, among the three IRT-based equating methods, the raw-to-raw conversions produced by the GRM equating methods diverged most from the population conversions produced by the equipercentile method. This finding is not consistent with Lee et al.'s (2001) finding that the GRM equating method produced results more consistent with those of the equipercentile method than the 3PL method. This inconsistency might be caused by the model selection difference: the 2PL was selected in this study whereas the 3PL was selected in Lee et al.'s study and LID has been shown to have large impact on the *c*-parameter estimation. Also, using GRM as an alternative to accommodate LID requires combining item scores into testlet scores, which suffers from the loss of item response pattern information and would cause inaccuracy in the final conversion. On the other hand, the 2PL and TRT equating methods were found to produce comparable results that were more consistent with the results of the equipercentile method across all four test conditions. This may be the case because the whole equating process involves multiple stages. In the item estimation stage of the equating process, the TRT model has been shown to yield more accurate item parameter estimates than the 2PL when LID are present (Bradlow et al., 1999; DeMars, 2006; Wainer et al., 2000). However, in the number-correct score equating stage, the TSEs and OSEs using the 2PL might agree more with the equipercentile equating results than the TSEs and OSEs using the TRT model because the TRT TSE and OSE methods equate the two forms only through the primary trait and integrate out the testlet-specific traits, whereas the 2PL and equipercentile methods do not distinguish these two traits.

Second, in terms of final raw-to-raw conversions, the 2PL equating method was quite robust to the violation of local item independence. As LID increased from low to high, the overall level of discrepancy between the 2PL equating conversion and the population equipercentile equating conversion remained stable with statistically insignificant fluctuations from test to test. Finally, the IRT TSE and OSE methods yielded similar equating results. This finding is consistent with

some previous studies (Han, Kolen, & Pohlmann, 1997; Lord & Wingersky, 1984) that indicated the IRT TSEs and OSEs yielded "indistinguishable results" (Lord & Wingersky, 1984, p. 453). Caution should be taken when generalizing these findings to other situations as these findings are limited to the specific conditions in this simulation study.

We acknowledge that we employed a random groups design in this study to avoid any influence of scale transformation on the final conversion table. In practice, the nonequivalent group with anchor test equating design is widely used and could greatly complicate the testlet-based test equating process. More factors will need to be considered if the nonequivalent group with anchor test equating design is applied, such as how to include testlet items and form a content and statistically representative anchor set, how to extend traditional scale transformation methods to the testlet-based tests, and so on. Only a few studies on these issues have been carried out so far. However, it should be noted that although this study might simplify the equating process by using a random groups design, the equating methods used in this study are also applicable to the testlet-based test equating under the nonequivalent group with anchor test design, and the equating results found in this study are informative as well.

We also acknowledge that we used different computer programs to conduct item calibration and number-correct score equating in this study. During the calibration phase, BILOG-MG was used for the 2PL, PARSCALE for the GRM, and SCORIGHT for the TRT model. The program SCORIGHT is able to calibrate all models in this study. We recommend using SCORIGHT for future studies as the use of a single program will avoid the possible confounding of model effects with estimation methods. We selected BILOG-MG for the 2PL and PARSCALE for the GRM for this study because they are commercially available and routinely used by practitioners in this field. During the equating phase, the program POLYEQUATE was used to conduct TSE and OSE with the 2PL and the GRM, and an SAS program was written for the TRT TSE and OSE. The program-dependent issues between the 2PL and the TRT model were minimal because the SAS program was also applied to the 2PL TSE and OSE and the same results were obtained from both POLYEQUATE and SAS. The specifications of GRM are completely different from the other two models, and thus, it is not possible to compare equating results across equating programs.

The TRT model, as a development from the traditional IRT models in the past few decades, provides more flexibility and accuracy to model testlet-based tests while retaining the item parameter interpretations as they are in the traditional IRT models. However, very little research has focused on testlet-based test equating using the TRT model or compared the equating results obtained under different IRT models in the presence of LID caused by testlets. Our study intended to fill in this research gap and found that the 2PL was adequate when the focus of the testing program was to generate raw-to-raw conversion tables. Given the prevalence of testlets and the popularity of applying the traditional dichotomous IRT models in practice, this finding has practical implications for test developers.

References

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153 – 168.

Cao, Y., Yin, P., & Gao, X. (2007, April). *Comparison of IRT and classical equating methods for tests consisting of polytomously-scored items*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement*, *43*, 145 – 168.

Gibbons, R. D. & Hedeker, D. R. (1992). Full-information bi-factor analysis. *Psychometrika*, *57*, 423 – 436.

Han, T., Kolen, M., & Pohlmann, J. (1997). A comparison among IRT true- and observed-score equatings and traditional equipercentile equating. *Applied Measurement in Education*, *10*, 105 – 121.

Keller, L., Swaminathan, H., & Sireci, S. G. (2003). Evaluating scoring procedures for context-dependent item sets. *Applied Measurement in Education*, *16*, 207 – 222.

Kolen, M. J. (2004). *POLYEQUATE* (*Windows Console version*) [Computer software and manual]. Iowa City: University of Iowa.

Kolen, M. J. & Brennan, R. L. (2004). *Test equating: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag.

Lawrence, I. M. (1995). *Estimating reliability for tests composed of item sets* (Research Report No. RR-95-18). Princeton, NJ: Educational Testing Service.

Lee, G., Kolen, M. J., Frisbie, D. A., & Ankenmann, R. D. (2001). Comparison of dichotomous and polytomous item response models in equating scores from tests composed of testlets. *Applied Psychological Measurement*, *25*, 357 – 372.

Li, Y., Bolt, D. M., & Fu, J. (2005). A test characteristic curve linking method for the testlet model. *Applied Psychological Measurement*, *29*, 340 – 356.

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement*, *30*, 3 – 21.

Lord, F. M. & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings.". *Applied Psychological Measurement*, *8*, 453 – 461.

Muraki, E., & Bock, R. D. (2003). *PARSCALE 4.1* [Computer software and manual]. Lincolnwood, IL: Scientific Software International.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*, 237 – 247.

Tang, K. L., & Eignor, D. R. (1997). *Concurrent calibration of dichotomously and polytomously scored TOEFL items using IRT models* (TOEFL Technical Report No. TOEFL-TR-13). Princeton, NJ: Education Testing Service.

Tao, W., & Cao, Y. (2012). *IRT true score and observed score equating with the testlet response theory model*. Manuscript submitted for publication.

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 245 – 269). Dordrecht, the Netherlands: Kluwer.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York, NY: Cambridge University Press.

Wainer, H. & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*, 185 – 201.

Wainer, H. & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement*, *37*, 203 – 220.

Wang, X., Bradlow, E. T., & Wainer, H. (2005). *User's guide for SCORIGHT* (*Version 3.0*): *A computer program for scoring tests built of testlets including a module for covariate analysis* (Research Report No. RR-04-49). Princeton, NJ: Educational Testing Service.

Zeng, L., Kolen, M. J., Hanson, B. A., Cui, Z., & Chien, Y. (2005). *RAGE-RGEQUATE Manual* (*Console version*) [Computer software and manual]. Iowa City: University of Iowa.

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2002). Identification and evaluation of local item dependencies in the Medical College Admissions Test. *Journal of Educational Measurement*, *39*, 291 – 309.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG 3* [Computer software and manual]. Lincolnwood, IL: Scientific Software International.

Zu, J. & Liu, J. (2010). Observed score equating using discrete and passage-based anchor items. *Journal of Educational Measurement*, *47*, 395 – 412.

Suggested citation:

Cao, Y., Lu, R., & Tao, W. (2014). *Effect of item response theory (IRT) model selection on testlet-based test equating* (ETS Research Report No. RR-14-19). Princeton, NJ: Educational Testing Service. doi: 10.1002/ets2.12017

Action Editor: Rebecca Zwick

Reviewers: Frank Rijmen and Jinghua Liu